Logo-Constrained Image Generation with ControlNet and IP-Adapter

Brage Dybesland Ramberg
Oslo Metropolitan University
brram4666@oslomet.no

Dung Thuy Vu
Oslo Metropolitan University
duvu35218@oslomet.no

Vaskar Shrestha
Oslo Metropolitan University
vashr0444@oslomet.no

***Abstract*—Recent advances in text-to-image diffusion models have enabled one to generate high-quality images from textual descriptions. However, such models still struggle with tasks that require fine spatial control and accurate reproduction of specific visual details, such as logos. This project explores a multimodal approach to automated banner creation with the combination of ControlNet and IP-Adapter to address these limitations. ControlNet provides spatial conditioning through inputs such as segmentation maps or edge maps, ensuring that design elements are accurately located. At the same time, IP-Adapter supports image prompt injection, allowing visual identity and style to be captured from seen images such as logos or brand icons. We develop a generation pipeline that processes pre-defined text prompts and images of logos and places the logos at specified positions without distortion. We demonstrate via comparative analysis that using spatial and visual conditioning works better than text-alone or single-modality generation for layout consistency and visual precision tasks. The findings highlight the benefits of combining structural and visual prompts in diffusion models for controllable, high-fidelity image generation, most notably in design-critical applications like marketing banners.**

***Index Terms*—Text-to-Image Generation, Diffusion Models, Stable Diffusion XL, ControlNet, IP-Adapter, Logo Placement, Visual Prompting, Image Synthesis, Generative AI, Masking Strategies, Grid Search, Banner Design**

I. Introduction

In recent years, Text-to-Image (T2I) generation using diffusion models has achieved remarkable progress, representing a paradigm shift in visual generation [1, 3, 7, 18, 20]. These models have demonstrated impressive capabilities in generating high-quality images from text prompts [11, 19]. Prominent examples include Google's Imagen [14, 20], OpenAI's DALL-E2 [17], and, notably, Stable Diffusion (SD) [21], which has been open-sourced and has showcased impressive capabilities in generating photorealistic images from text prompts. Diffusion models work by gradually denoising a random state into a clear image, and T2I diffusion models condition this denoising process on text inputs [1, 5, 20].

The impressive capabilities of T2I diffusion models have opened up new possibilities for creative expression, design, and the generation of multimedia content [1, 2, 7]. Text is highlighted as an intuitive and powerful way to condition image generation, making content creation accessible even to individuals without specialized expertise [5, 20]. As generative models continue to reshape creative industries, one practical area that has attracted attention is the automatic generation of marketing visuals, such as banners, posters, and advertisements [1, 3].

This project explores the specific challenges in generating banners with predefined prompts and fixed logo placements without compromising visual quality or introducing distortions. Despite the remarkable progress of text-to-image diffusion models, relying only on textual input often proves inadequate in tasks that require spatial layout and detailed visual control [3, 10]. Text prompts lack the expressiveness to accurately convey complex design layouts, making it difficult to position elements at specific locations or to maintain consistent style and composition [3, 5]. These models usually struggle with rendering legible text, accurately replicating unseen logos, or preserving detailed structural relationships, often leading to misalignment, distortion, or unintended output. Moreover, text-based conditioning provides only coarse control over spatial arrangements, as it cannot effectively encode positional data or dense layout information [10]. As a result, generative models tend to hallucinate or deform small but critical elements, such as logos, when trying to align both textual and spatial cues in a complex visual structure. This underscores the need for multimodal conditioning methods, such as ControlNet and IP-Adapter, which enable structure-aware and image-guided generation to meet the precise requirements of banner creation.

To address this challenge, we leverage the capabilities of Stable Diffusion augmented with two recent advances: ControlNet and IP-Adapter. ControlNet enables fine-grained control over image generation by conditioning the diffusion process on additional structural inputs such as edge maps, pose annotations, or segmentation masks [10, 19]. The IP-Adapter, on the other hand, allows image prompts to be injected into the generation process, enabling visual conditioning and using reference images such as logos [17]. When combined, these tools allow us to control both the content and spatial composition of the generated images more reliably than text prompts alone [17].

The primary goal of this project is to generate banners that (1) follow a predefined text prompt, (2) incorporate a specific logo, and (3) ensure that the logo appears exactly at one of the pre-defined positions. Importantly, the generated logos should be free of distortion and visually coherent with the rest of the banner. We also also aim to evaluate how well the models preserve visual fidelity while adhering to strict layout constraints, which are crucial for branding and marketing applications.

Our approach combines prompt engineering with visual control techniques to construct a pipeline capable of generating high-quality banners under these constraints. We explore different methods of integrating ControlNet and IP-Adapter into the generation process, analyzing the results both qualitatively and quantitatively.

II. Literature Review

A. Introduction to Diffusion-Based Text-to-Image Generation

The field of image generation has progressed significantly from earlier methods like VAEs and Generative Adversarial Networks (GANs) to today's powerful diffusion-based approaches [1, 3, 20]. Although GAN-based text-to-image generation showed early promise, it often had limitations. Some of them were that the model being confined to domain-specific datasets, generating images with fewer parameters, struggled to generate complex or detailed images, and capturing zero-shot text descriptions compared to later diffusion models [1]. In contrast, diffusion models introduced a paradigm shift in the way generative models synthesize images. Taking inspiration from thermodynamics, their concept involves a forward process in which noise is gradually added to an image until it becomes completely degraded into noise [3, 18]. The fundamental idea is to train a deep neural network to learn the reverse process, which is the iterative removal of this noise, ultimately revealing a clear image [20].

Text-to-Image (T2I) diffusion models extend this fundamental denoising process by making it conditional on textual input via encoders such as CLIP or Text-To-Text Transfer Transformer (T5) [6, 3, 18]. In this process, image generation is guided by a user-provided text prompt, which is converted into meaningful latent representations by a text encoder [9, 21]. The cross-attention layers within the diffusion model architecture are implemented within a U-Net structure that allows the models to align the text features with the generation of images, supporting the production of coherent and comprehensive responses [1]. U-Net is a convolutional neural network with both an encoder and a decoder structure. However, as the denoising progresses toward the targeted image, the model gradually shifts focus from textual conditioning to enhancing visual quality.

Several key models have emerged, showcasing remarkable capabilities and the power of T2I diffusion models in terms of image quality and diversity. Google's Imagen has achieved an unprecedented degree of photorealism and advanced language understanding, where they found that the use of large, frozen language models as encoders can significantly improve the quality of the image and its alignment [14]. OpenAI's DALL-E 2 introduced a two-stage approach combining CLIP-based image embedding on the text caption, and a diffusion-based decoding, that supported both text and image prompts [4, 18]. Meanwhile, another latent diffusion model, Stable Diffusion, has become a prominent example due to its open-source accessibility, which has led to its widespread usage [3, 19, 21]. Together, these developments have achieved impressive success in creative generation through their ability to generate high-fidelity images that accurately reflect descriptions in natural language.

B. Limitations of Text-Only Conditioning

While text-to-image diffusion models have made remarkable advancements in generating high-quality images, only depending on textual prompts presents significant limitations, especially in tasks that demand precise control over visual attributes or nuanced customization [3, 10, 22]. Text prompts are often inadequate for expressing complex concepts, such as unfamiliar logos, specific art styles, or detailed object layouts. Diffusion models struggle to represent structural information like object position, orientation, and relationships through text alone, making it difficult to generate images that align with specific spatial layouts or constraints [3, 5, 17]. In addition, text-guided synthesis, particularly with complex text involving multiple subjects or enriched descriptions, often encounters issues of textual misalignment [3].

Moreover, rendering legible and stylistically accurate text within images remains a largely unsolved challenge [21]. Standard text encoders such as CLIP fail to sufficiently capture character-specific information for text rendering, leading to poor fidelity in text rendering and limited variation in styles or fonts. Similarly, precise logo reproduction is difficult because of the unique patterns and textual components logos often contain, especially when such logos are not part of the model's training data [22]. Diffusion models frequently misplace or distort logos, as they cannot interpret or replicate fine-grained relationships between image elements using text instructions alone [20, 22].

Empirical studies that used benchmarks like DrawBench highlight persistent issues such as spatial misalignment, missing objects, and corrupted logo identity when only text is used for conditioning [14, 20]. Qualitative comparisons of models attempting tasks like inserting logos demonstrate that methods relying primarily on text-based approaches or without specific training for logo insertion often struggle to preserve the logo identity accurately [22]. These challenges usually require extensive trial-and-error with prompts, and even then, do not guarantee desired outcomes. These findings underscore the necessity of integrating additional forms of conditioning, such as spatial maps or image prompts, to enable more accurate, consistent, and controllable image generation.

C. Multimodal and Spatial Conditioning Methods

To address the limitations of text-only conditioning in diffusion models, particularly in tasks requiring fine-grained control and accurate logo placement, recent research has introduced multimodal conditioning methods that go beyond text. These methods fall into three main categories: spatial maps, image prompts, and layout-based controls [3, 19, 17, 20].

Spatial maps provide structural information such as object positions and dense labels, enabling more precise spatial control than text alone [3, 19]. A prominent neural network architecture is ControlNet, which adds spatial conditioning controls, such as edge maps, depth maps, segmentation maps, human pose skeletons, and normal maps, to large, pretrained text-to-image diffusion models [19]. ControlNet works by locking the original pretrained diffusion model and adding a parallel network that copies the original model's encoder layers [19]. It can operate with single or multiple conditions and can even generate images without any text prompt, robustly interpreting content semantics from the conditioning image. ControlNet's simplicity and flexibility have made it widely adopted for tasks requiring layout awareness.

Beyond structural maps, incorporating reference images as conditions allows for transferring visual information, such as style or specific content. IP-Adapter, an image prompt method, is designed to provide prompt capability for pre-trained text-to-image diffusion models without modifying the original text-to-image models [17]. Unlike earlier methods that replaced text encoders, IP-Adapter uses a dual attention mechanism to support both image and text prompts simultaneously by separating cross-attention layers for text features and image features [3, 17]. This design enables the conditioning of diffusion models on both style and content, making it especially suitable for inserting complex logos that cannot be effectively described with text alone. While textual customization methods like DreamBooth and Textual Inversion have been explored, they often fail when models lack prior knowledge of the logo or visual element [22].

While spatial maps provide dense, detailed control, coarser spatial information like bounding boxes and layout tokens can also be used to guide generation. This approach provides a form of spatial guidance, helping to organize multiple objects in structured compositions [3, 10, 20]. Layout information, such as the positioning of objects, has demonstrated notable improvements in generating multiple objects in image synthesis. Nonetheless, a key challenge lies in the generation of the layout information itself.

Together, these advanced conditioning techniques provide varying degrees of spatial control, moving beyond the limitations of text alone, enabling more accurate and coherent generation of structured visuals like banners with embedded logos.

D. ControlNet: Structure-Aware Control

ControlNet works by locking certain layers of the original pretrained diffusion model to keep its strong generative capabilities intact [19]. It then duplicates the original model's encoder layers into a trainable parallel network, commonly known as the "control model". This model processes the conditioning input, such as an edge map, through additional convolutional layers, including using "zero convolutions" to effectively encode structural guidance [3, 9]. These layers begin with zero weights to ensure the original model remains unaffected at the start of training [19]. The control model's output is fused with the main model's features before upsampling, allowing the spatial signal to guide image synthesis without disrupting the base model's generative capacity [9]. As a result, ControlNet becomes both effective and compatible with a wide range of tasks requiring fine control.

ControlNet can operate with single or multiple conditions, and can function with or without accompanying text prompts. For banner design, using controls like segmentation maps or human pose can be invaluable for controlling the positioning of elements like logos, product images, or background sections, enabling precision in layout design that text prompts cannot achieve [19, 9].

E. IP-Adapter: Image Prompt Injection

Complementing ControlNet's spatial guidance, IP-Adapter introduces a method for conditioning diffusion models on reference images. IP-Adapter is a lightweight module that augments a pre-trained text-to-image model by injecting a second stream of attention based on image prompts [17]. The key design of IP-Adapter is a decoupled cross-attention mechanism that separates the cross-attention layers used for text features from those used for image features. It introduces an independent cross-attention layer within the UNet architecture, decoupled from the text-based attention. The image prompt is encoded using a CLIP image encoder, projected through a small trainable network, and then fed into these new attention layers. In the training process, only the adapter components are trained while the pretrained diffusion model, the original UNet, is kept frozen [17]. Only the newly added cross-attention layers and a projection network are trained. IP-Adapter is trained on a multimodal dataset of text-image pairs, is lightweight with a small number of trainable parameters. The system works with both image-only and combined image-text prompts, giving users flexibility in conditioning generation.

F. Comparative Analysis and Hybrid Approach

Achieving precise spatial control and accurately reproducing specific visual identities, such as logos or brand elements, remains a challenge for text-to-image diffusion models. While text prompts alone are often insufficient for conveying detailed layouts or specific visual attributes, there has been little to no any research study on the performance of integrating additional conditioning methods. Comparative studies show that combining spatial conditioning (ControlNet) with visual identity injection (e.g., IP-Adapter) leads to significantly better performance than using either method alone.

By combining ControlNet for layout and IP-Adapter for visual identity, users can achieve highly controlled image synthesis. For instance, in banner generation, ControlNet can define the placement of a logo via segmentation maps, while IP-Adapter ensures the logo's visual appearance remains accurate through reference image conditioning. This will controls what to generate and where to generate during image generation.

In banner design, where both precise spatial layout and faithful reproduction of logos or product visuals are essential, combining ControlNet and IP-Adapter proves highly effective. This hybrid approach ensures visual elements are correctly placed and visually accurate, meeting the high demands of marketing and brand consistency.

III. Proposed solution

A. Approach Overview

In this project, we propose a multi-modal framework integrating Stable Diffusion XL with ControlNet for spatial guidance and IP-Adapter for visual conditioning. SDXL provides enhanced semantic understanding through dual text encoders (CLIP ViT-L and OpenCLIP ViT-bigG). ControlNet ensures precise spatial positioning via structural conditioning, while IP-Adapter preserves brand identity through CLIP-based visual encoding. This architecture enables independent control over placement and fidelity.

B. Pipeline Framework

Our pipeline comprises: (1) dual-modal encoding, (2) multi-modal conditioning, (3) temporally-modulated generation, and (4) qualitative output inspection. Text prompts undergo dual encoding while logos are processed through CLIP ViT-H. Generation integrates both conditioning streams via decoupled cross-attention with hard cutoff strategies to prevent concept bleeding.

C. Latent Diffusion Models and Stable Diffusion Architecture

Stable Diffusion addresses computational challenges by operating in compressed latent space rather than pixel space through Variational Autoencoder (VAE) compression. The VAE encoder compresses input images $x \in \mathbb{R}^{H \times W \times 3}$ into latent tensors $z \in \mathbb{R}^{h \times w \times c}$ with $8\times$ spatial reduction and 4 channels, providing computational efficiency and perceptual compression. The denoising process employs a UNet architecture with encoder-decoder structure, ResNet-style blocks, and skip connections for spatial information preservation. Cross-attention mechanisms within Transformer blocks enable text-conditional generation by integrating semantic information from language models into the denoising process. This VAE-latent compression combined with Transformer-enhanced UNet enables efficient, high-quality conditional image generation with reduced computational requirements.

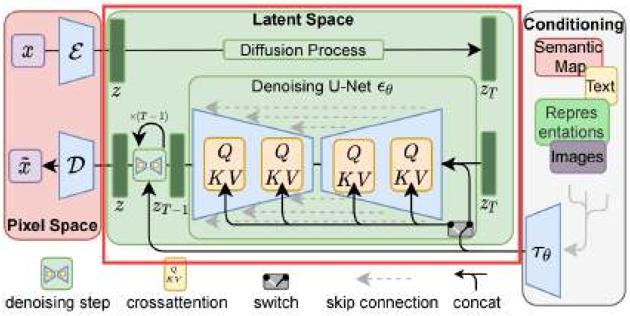


Fig. 1: UNET architecture of LDM

D. Stable Diffusion XL (SDXL) 1.0: Architectural Advancements

Stable Diffusion XL (SDXL) 1.0 represents a substantial architectural refinement over previous latent diffusion models, incorporating enhanced conditioning mechanisms to improve text-image alignment and high-resolution synthesis capabilities [12]. The architecture integrates dual text encoder systems with micro-conditioning mechanisms that enhance semantic comprehension and compositional controllability in generated outputs. SDXL implements size and crop conditioning through Fourier feature embeddings, encoding spatial metadata directly into the generation process for precise control over image resolution and layout composition. Size conditioning utilizes Fourier embeddings $\gamma(x) = [\sin(2^i \pi x), \cos(2^i \pi x)]$ with $L = 12$ frequency bands, generating 24-dimensional vectors integrated with timestep embeddings to condition output resolution. Crop conditioning employs identical Fourier encoding for spatial coordinates $\gamma_{\text{crop}}(x_{\text{crop}}, y_{\text{crop}}) = [\gamma(x_{\text{crop}}), \gamma(y_{\text{crop}})]$, enabling precise regional control essential for spatial manipulation tasks. The crop mechanism facilitates controlled generation in designated regions, particularly valuable for applications requiring precise logo positioning within specific spatial boundaries. Both conditioning signals are concatenated and integrated with timestep representations, enabling dynamic adaptation of generation trajectories based on spatial and compositional constraints. This micro-conditioning framework ensures layout preservation while accommodating spatial requirements, providing enhanced control for logo-constrained generation applications.

E. Dual Text Encoder Architecture

SDXL employs a dual text encoder architecture combining CLIP ViT-L (12 layers, 768-dimensional, 123M parameters) and OpenCLIP ViT-bigG (24 layers, 1280-dimensional, 694M parameters) to enhance semantic comprehension beyond single-encoder limitations. The encoders process prompts in parallel, with concatenated embeddings forming $E_{\text{combined}} = [E_{\text{CLIP}}(x); E_{\text{OpenCLIP}}(x)]$ (2048 dimensions) and pooled global embedding $E_{\text{pooled}} = \text{Pool}(E_{\text{OpenCLIP}}(x))$ integrated with timestep embeddings. This dual-path conditioning ensures both local token granularity and high-level linguistic structure are embedded in the generative process, providing superior semantic understanding and text-image alignment.

F. IP-Adapter for Reference Image Guidance

IP-Adapter introduces visual reference conditioning to complement SDXL's text-based control and ControlNet's spatial guidance, enabling the incorporation of reference images into the generation pipeline for applications where textual descriptions alone cannot capture specific stylistic or structural constraints such as logo placement and visual identity preservation [16].

1) Image Feature Extraction: The encoder generates patch-wise embeddings in 2D spatial format where each token corresponds to a specific image region, subsequently projected through a learnable layer into UNet-compatible space:

$f_{\text{img}} = \text{Proj}(\text{CLIP}(x_{\text{ref}})) \in \mathbb{R}^{257\times 1024}$. This projection maps high-dimensional CLIP features to the 1024-dimensional space expected by UNet's attention layers, providing dense embeddings that encode both compositional structure and aesthetic elements for precise visual conditioning.

2) Decoupled Cross-Attention Mechanism: IP-Adapter implements a decoupled cross-attention mechanism that computes separate attention distributions over text and image key-value pairs to maintain modality-specific representational integrity during joint conditioning:

$$\text{Output} = \text{softmax}\left(\frac{QK_{\text{text}}^T}{\sqrt{d}}\right)V_{\text{text}} + \lambda\cdot\text{softmax}\left(\frac{QK_{\text{img}}^T}{\sqrt{d}}\right)V_{\text{img}} \tag{1}$$

where Q denotes query matrices from UNet hidden states, K and V represent keys and values from respective modalities, and scalar parameter λ regulates image guidance contribution. This architecture enables task-specific modality weighting while preventing visual conditioning from dominating textual semantics. The projection layer maps CLIP features to 1024-dimensional UNet-compatible representations, generating feature matrix f_{img} that encodes structural and aesthetic properties for precise visual conditioning.

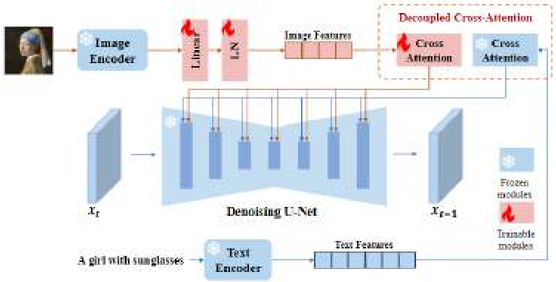


Fig. 2: IP Adapter architecture

G. Data Preparation

The initial preparation involved selecting logos from the publicly available *Logo Images Dataset* on Kaggle [13]. Although this dataset offered broad variety, we found it unnecessary to use a large number of logos and they were also low in resolution.

Instead, we shifted our focus to high-definition (HD) logos, sourced from Logo Ipsum [8]. These HD logos—especially in PNG format with transparent backgrounds—proved significantly more effective for visual conditioning via IP-Adapter due to their clarity and clean shape structure.

H. Visual Interpretation

The primary method for evaluating model performance throughout the project was visual inspection. All tuning of parameters were done by assessing the aesthetic quality and logo preservation in the generated images. This hands-on, qualitative approach allowed for rapid iteration and identification of effective configurations.

Although objective evaluation metrics exist for generative models, we realized that this would be very difficult to adapt to our specific case. Such metrics include CLIP score, FID (Fréchet Inception Distance), or SSIM (Structural Similarity Index). Attempting to score both global coherence and localized logo accuracy mattered. Ideally, we considered designing a custom metric that would evaluate both the overall visual appeal of the image and a cropped region around the logo to assess fidelity and aesthetics. Building such a pipeline would have required substantial additional effort and GPU resources, which was beyond the scope of this project. As a result, we relied on qualitative visual interpretation as our evaluation method.

IV. Experiments

A. Compute Resources and Infrastructure

During early development, we attempted to run our generation pipeline locally on personal laptops. However, we quickly encountered significant limitations in both memory and processing power. These constraints made it clear that a more powerful environment was necessary to iterate efficiently and conduct meaningful experimentation.

To address this, we migrated our workflows to the eX3 infrastructure hosted by SIMULA Research Laboratory. The research presented in this paper has benefited from the Experimental Infrastructure for Exploration of Exascale Computing (eX3) [15].

Specifically, we utilized compute nodes equipped with high-performance NVIDIA GPUs, including A100, HGX, and DQX2 systems. These hardware specs are capable of handling large-scale inference workloads and multi-image grid searches. These GPUs, with their large memory capacity and support for mixed-precision acceleration, were essential for enabling large-scale inference, multi-parameter grid searches, and efficient experimentation with ControlNet and IP-Adapter pipelines.

1) Grid Search and parameters: To explore the interaction between the parameters, we conducted multiple extensive grid searches over key parameters:

- **G (guidance scale)**: Controls how strongly the prompt guides generation. Typical range: **[5.0 – 9.0]**. Lower values increase image diversity; higher values enforce prompt fidelity.
- **A (adapter scale)**: Strength of the IP-Adapter's influence on generation. Range: **[0.0 – 1.0]**. Values above 0.6 often caused distortion; sweet spot was found around 0.1–0.3.
- **CN (ControlNet scale)**: Controls the impact of the ControlNet conditioning input. Range: **[0.0 – 1.0]**. Best results were found between 0.7 and 1.0.
- **CO (cutoff step percentage)**: Adapter influence is turned off after this percentage of the total steps. Given as an integer percentage, e.g., `CO20` = 20% of steps. Helps avoid overfitting to the logo. Common values: **10 - 33**.
- **SS (steps)**: Number of denoising steps in the diffusion process. Common values: **50–100**, but we narrowed it

down to **50** due to low returns, quicker compute, and the risk of logo overwriting at high step counts.

The goal was to identify combinations that yielded high-quality generations while preserving the logo accurately in the designated region. We systematically varied each parameter and generated over large image samples, starting of with many parameters varying between intervals described in the list above.

As described in the Visual Interpretation subsection, the primary evaluation metric was manual visual inspection of the results, focusing on logo clarity, placement consistency, and overall image realism. The grid search not only refined our parameter ranges but also confirmed our assumptions that early suppression of the adapter signal (via cutoff) and moderate adapter strength were critical to achieving both logo fidelity and coherent scene generation.

B. Prompting

Including the logo color in one prompt example overrode the IP-adapter, something proving great control over the generation. Mountain example Figure 3 with white logo, shows the fox logo in white, and the generated image has a ”white” theme based on visual inspection. Given the similar prompt, but with the color specified as white, it created the same picture but with the original orange logo color instead. This one has a sunset feel to it, matching the color in the original logo.

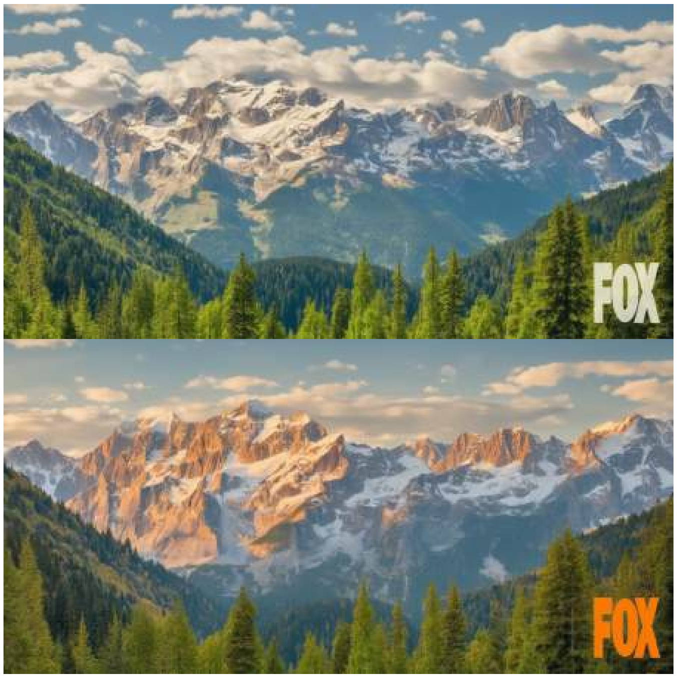


Fig. 3: Generated images with white logo (top) and orange logo (bottom) variants. Both with same parameters: g=6.0, A=0.2, CN=0.7, CO=20, SS=50

The only change between the two prompts was the removal of the word **white**:

```
"A high-resolution scenic mountain
banner, majestic snow-capped peaks under a
clear blue sky, lush green forest at the
base, crisp atmosphere. There is a white
logo in the bottom-right corner"
```

This shows high control with prompt engineering, but this specific feature was not explored further due to limited time. Other smaller experiments were tried on this matter, and other results showed distortion or unwanted artifacts when trying to add descriptions of the logo. However, assumptions like these need to be inexpressibly tested with focus primarily on the guidance scale, which is not the main goal for our project.

C. Image Masks

Early on the canny edge mask for ControlNet proved great to let the diffusion process not influence the logo freely, allowing control with higher step counts. Experimentation showed that high step counts often overrode the logo which is why the canny edges were present through the model iterations.

Further into development, we experimented with multiple masking techniques to guide the generation specific to the IP-Adapter. Initially, this mask was created by blurring the region corresponding to the logo placement attempting to focus the attention on the adapter. The hypothesis was that this soft mask should give smooth transitions, it often introduced undesirable artifacts, most notably a glowing effect around the logo and in some cases completely distorted the logo. This glow interfered with background details and blended poorly with the scene.

After further experimentation, we settled on a split strategy: using the *Canny edge mask*for the ControlNet input, and a *hard alpha mask* for the IP-Adapter. The Canny mask emphasizes the logo’s shape in a way that guided the diffusion process to respect its form. Meanwhile, the alpha mask provides a precise and bounded region for the IP-Adapter to influence, preserving sharpness and minimizing interference with the surroundings. This combination proved most stable and visually aesthetic.

1) Dynamic IP-Adapter Scaling: Given the way stable diffusion work we realized that having high adapter values throughout the generation was never going to guide the image into a nice banner, when the input for the adapter were just a logo. As seen in Figure 4, the image are highly distorted and notably the image is highly affected by concept bleed. Grid search proved that adapter values in early stages of the iteration were helpful, guiding the logo generation early and cutting them off once they ”set”. For this we applied two strategies. Dynamic adapter scaling, where after a certain amount of steps the IP-Adapter scale values were slowly turned off.

This approach however seemed to be highly sensitive to all the parameters given, and therefore without having a good metric to set up for random search, we ended up with a hard cutoff strategy instead. The hard cutoff were easily implemented and given as a parameter to the pipeline. After a given step, the IP-Adapter turns completely off, and will let ControlNet and the rest of the generation take over.

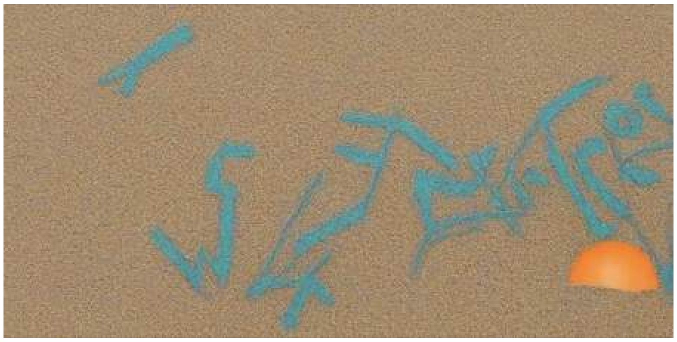


Fig. 4: Distorted by consequent high adapter scale. Parameters: G=4.0, A=0.6, CN=1.0, CO=100, SS=50

V. Results

For the qualitative observations we aim to focus on mainly IP-Adapter strength and cutoff scheduling. To evaluate the effect they have on image generation quality and logo preservation, we present grid collages in Appendix VIII. Each grid show the base prompt and varies IP-Adapter scale (A) along rows and cutoff percentage (CO) along columns.

A. Impact of IP-Adapter Strength (A)

Across grids, increasing the adapter strength generally led to better reproduction of the logo in areas like clarity, color fidelity, and sharpness. However, at higher strengths particularly in conjunction with high cutoff values signs of over-conditioning emerged for difficult input logo shown in Figure 11 and 12.

Visual inspection shows overly dominant logo features, such as overcompensation of coloring, and concept bleed. Additionally, som results showed loss of global coherence, where background elements appear distorted or washed out. Another aspect of the IP-adapter is also that higher values and more complex logos will affect the image generated as a whole, something that means that the IP-Adapter is affecting the stable diffusion process. This theory is derived by comparing Figure 11 and 12 with the other Beach Scene Figure 10, with the notable difference that the logo is different input.

The less difficult ”FOX” logo Figure 6, shows overall good fidelity both in Figure 7 and 10. Higher adapter specifically in the Forest Scene show that weird artifacts are better handled, as visual inspection shows something similar to an umbrella in low Adapter and low cutoff parameters. One hypothesis is that the less noise and features from the input, the logo will be more dominant and and easier guided in the image generation.

B. Effect of Cutoff Percentage (CO)

The cutoff parameter controls how long the IP-Adapter influences the generation process. Lower cutoffs tend to preserve scene realism while allowing the logo to settle early without bleeding as much into the entire image. On the other hand, the small window for the adapter to enact may lead to distortion of the logo as the rest of the diffusion process takes over.

Higher cutoffs often result in excessive logo influence, like color shifts, especially in low-texture regions like water, sky or sand. Figure 12 shows a clear bleed of color in almost all high values for adapter and cutoff. A tradeoff can be seen where high adapter does recreate the logo better in some areas (excluding the pink blob of color on top of logo itself).

C. Scene Complexity and Prompt Sensitivity

The structural complexity of the background scene played a significant role in how logo distortions manifested. Scenes with rich textures and depth, such as forests, were more forgiving, allowing the model to embed logos with fewer visible artifacts. In contrast, low-texture environments tended to amplify bleeding effects from the adapter, especially under high cutoff values. This suggests that controlling for background context may be important when generalizing this pipeline to diverse use cases.

Although the visual conditioning was primarily handled via IP-Adapter and ControlNet, we observed that prompt phrasing could significantly influence results. For instance, explicitly including color cues (e.g., “purple logo”) sometimes caused global hue shifts in the generated image, leading to unnatural blending or a visual echo around the logo. This suggests that textual conditioning remains active throughout generation and may reinforce or interfere with adapter signals.

Specifically the Mountain Scene Figures 8 and 8, had some unexpected outcomes. Mentioning the purple logo in the prompt resulted in better logo influence on the scene, however this resulted also in a purple circle around the logo itself. Attempting to remove the purple logo part from the prompt resulted in white transparent logo instead, however, remaining in shape towards the input.

D. General Trends

- **Low Adapter + Low CO**
  Minimal logo impact. The logo is often underdeveloped, faint, or blurry, with minimal influence on the overall image.
- **Mid Adapter + Mid CO**
  Provides the best balance between visual realism and logo integrity. The logo is clear and well-placed, and the scene remains coherent and visually appealing.
- **High Adapter + High CO**
  The logo becomes overly dominant, often leading to distortion of background elements. This can result in flat textures, color shifts, or loss of scene fidelity.

VI. Discussion

A. Advancing Logo-Constrained Generation: From Structural to Multi-Modal Conditioning

1) From Single-Modal to Multi-Modal Conditioning: Initial ControlNet experiments using spatial conditioning (edge detection and position masking) enabled logo placement but failed to preserve visual identity. This shortcoming prompted a shift to multi-modal approaches that combine spatial and semantic control for improved fidelity. Relying solely on structural conditioning caused significant degradation in logo fidelity and background quality, making it unsuitable for commercial applications due to loss of critical brand features such as color, texture, and shape.

2) Comparative Workflow Analysis: The ControlNet-only pipeline treated logos as structural outlines via edge maps and fixed masks. Though effective for positioning, it generated synthetic data with prompt misalignment, visual artifacts, and poor contextual integration, limiting practical use. Introducing IP-Adapter added dual-conditioning: spatial guidance from ControlNet and visual identity preservation via CLIP embeddings. This hybrid design addressed fidelity issues and enabled precise logo placement. Temporal modulation techniques, such as hard cutoff scaling, further enhanced control and reduced concept bleeding.

3) Limitations of Single-Modal Conditioning: The ControlNet-only method suffered from artifacts and inconsistent prompt-output alignment due to synthetic inpainting data. Edge-based conditioning lost critical logo details like color and typography, reducing logos to mere outlines and undermining brand fidelity. Scalability was limited due to poor generalization across diverse real-world cases.

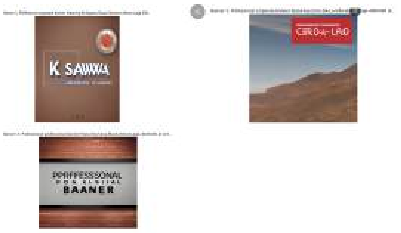


Fig. 5: Illustrative results of banner generation using a single-modal approach

4) Benefits of Multi-Modal Conditioning: Combining ControlNet with IP-Adapter overcame these limitations by enabling simultaneous spatial control and visual identity preservation. CLIP-based encoding maintained logo-specific features, enhancing fidelity and brand consistency. Adaptive scaling and temporal modulation improved layout flexibility and mitigated concept drift, allowing fine-grained control over form and placement. This multi-modal framework showed clear improvements over single-modal methods and demonstrates promising potential toward meeting commercial design quality standards.

B. Limitations of the Proposed Multi-Modal Conditioning Approach

A key limitation is the lack of quantitative evaluation; reliance on visual inspection reduces scientific rigor and reproducibility of parameter settings. Experiments were limited to high-resolution PNG logos with transparent backgrounds, raising concerns about applicability to low-resolution, vector, or complex logos. Without failure analysis or robustness testing, practical deployment remains uncertain. Additionally, the approach assumes logos as fixed entities, preventing pixel-level contextual modifications, thus limiting flexibility for adaptive branding needs.

C. Future Implementation

1) Comprehensive Evaluation Framework: Future work should adopt quantitative metrics covering logo fidelity (e.g., SSIM, color difference), spatial accuracy (e.g., IoU, center offset), and perceptual quality (e.g., FID, CLIP alignment, LPIPS) across varied generation scenarios. Statistical rigor requires factorial designs, bootstrap sampling, ANOVA, and confidence intervals to validate parameter effects. Benchmarking should include comparisons with text-only baselines, individual modules, and commercial tools. A diverse dataset of 500+ logos from multiple industries, annotated with standardized ground truths and automated scoring, is essential.

With a metric for evaluation random search could be developed and a general pipeline could be presented with better tuning, and understanding on how each parameter influence the image generation. This could essentially lead to even better results based on real measurements, not just visual inspection.

2) Advanced Logo Control Architecture: Next-generation models should develop hierarchical conditioning to separate brand identity from presentation style, enabling semantic understanding, structural abstraction, and style encoding for flexible yet consistent logo modifications. Extending IP-Adapter to support localized image-to-image transformations will allow contextual logo integration via attention masking, progressive refinement, and constraint-aware optimization. Multi-scale logo processing should dynamically adjust conditioning based on logo and background characteristics, reducing manual parameter tuning.

3) Robustness and Production Deployment: Dynamic, content-aware parameter adaptation informed by real-time analysis (e.g., generation progress, logo complexity) can optimize adapter scaling and cutoff thresholds. Monitoring intermediate outputs can guide real-time adjustments to enhance consistency. Scalability demands progressive generation methods, optimized attention mechanisms, and model compression. A complete deployment pipeline should include automated preprocessing, intelligent placement, quality assurance, and iterative user feedback loops. Validation must encompass professional designer evaluations, A/B testing in commercial campaigns, blind comparisons, and reproducibility through shared datasets and open evaluation scripts.

4) Beyond Logos: Broader Applications of the Pipeline: Although this project targets logo placement in banner generation, the core pipeline can be adapted and improved to solve other real-world design problems. The ability to guide both visual style and element positioning makes it relevant for modern applications such as automated content creation, personalized marketing assets, or dynamic ad generation. Multimodal generative tools can become more integrated into commercial design workflows, and pipelines like this offer a practical step toward more controllable and efficient visual content generation.

VII. Conclusion

In this research study, we briefly explored the challenges that image-to-text diffusion models come across during image synthesis. Furthermore, we also studied the possible solutions in generating marketing banners using text-to-image diffusion models with both predefined textual prompts and specific logo placements. The study shows that diffusion models like Stable diffusion models have introduced a paradigm shift in image synthesis, but they still struggle with tasks requiring fine-grained spatial control and fidelity during the reproduction of complex visual elements like logos. Depending solely on textual prompts can often result in image distortions, misalignments, or the loss of visual identity, challenges that are very critical in designing digital marketing banners.

To overcome these challenges, our work combines integrating two powerful conditioning methods: ControlNet and IP-Adapter. ControlNet introduced spatial control into the generation process by accepting conditioning inputs like edge maps or segmentation masks so that objects like logos or text blocks are synthesized in predefined positions. Whereas, IP-Adapter introduces image prompt injection, allowing reference images, such as logos, to guide the model to preserve visual identity, texture during image synthesis. The combination of these two approaches becomes a hybrid pipeline that enhances both content accuracy and layout fidelity.

In this study, we tested this hybrid pipeline on a set of banner generations that included various pre-defined prompts and logos. Our results show that the combination of spatial and visual conditioning produces better results than text-only or single-modality methods. Upon visual inspections of the generated output, we found that multimodal guidance significantly improved the quality, consistency, and controllability of the generated images, highlighting the growing importance of multimodal conditioning in creative AI digital marketing banners.

Acknowledgment

We would like to express our sincere gratitude to everyone who supported us throughout this project. Special thanks to Professor Raju Shrestha, who guided us with his valuable feedback, and we would also like to thank our classmates who gave us constructive feedback during the presentation of our project. We would also like to acknowledge the use of AI-based tools, including ChatGPT, Notebook LLM, which were utilized to assist in various phases of the project. These tools supported us in evaluating and refining source code, improving the clarity and structure of our writing, and helping us brainstorm and articulate ideas. We would also like to express our sincere gratitude to SIMULA Research Laboratory [15] for providing access to their high-performance computing infrastructure. The availability of powerful GPU resources and support through the SIMULA eX3 cluster was essential for running our experiments efficiently and at scale. Their contribution made it possible to explore and evaluate the multimodal generation techniques used in this project.

Source Code:

The main implementation code is provided as a ZIP attachment with this report, including setup instructions and a detailed README. Code for prior approaches, including the single-model spatial conditioning method, is available in the accompanying Google Drive folder. Project Files on Google Drive

References

[1] Fengxiang Bie et al. “Renaissance: A survey into ai text-to-image generation in the era of large model”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[2] Rumeysa Bodur, Binod Bhattarai, and Tae-Kyun Kim. “Prompt augmentation for self-supervised text-guided image manipulation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 8829–8838.

[3] Pu Cao et al. “Controllable generation with text-to-image diffusion models: A survey”. In: *arXiv preprint arXiv:2403.04279* (2024).

[4] Zijie Chen et al. “Tailored visions: Enhancing text-to-image generation with personalized prompt rewriting”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 7727–7736.

[5] Guillaume Couairon et al. “Zero-shot spatial layout conditioning for text-to-image diffusion models”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 2174–2183.

[6] Hexiang Hu et al. “Instruct-imagen: Image generation with multi-modal instruction”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 4754–4763.

[7] Youwei Liang et al. “Rich human feedback for text-to-image generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 19401–19411.

[8] *Logo Ipsum - Free Logo Placeholders*. Accessed May 2025. URL: https://logoipsum.com/.

[9] Denis Lukovnikov and Asja Fischer. “Layout-to-Image Generation with Localized Descriptions using ControlNet with Cross-Attention Control”. In: *arXiv preprint arXiv:2402.13404* (2024).

[10] Sicheng Mo et al. “Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 7465–7475.

[11] Maitreya Patel et al. “Eclipse: A resource-efficient text-to-image prior for image generations”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 9069–9078.

[12] Daniel Podell et al. “SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis”. In: *arXiv preprint arXiv:2307.01952* (2023).

[13] Siddharth Sah. *Logo Images Dataset*. Accessed May 2025. 2021. URL: https://www.kaggle.com/datasets/siddharthkumarsah/logo-dataset-2341-classes-and-167140-images.

[14] Chitwan Saharia et al. "Photorealistic text-to-image diffusion models with deep language understanding". In: *Advances in neural information processing systems* 35 (2022), pp. 36479–36494.

[15] Simula Research Laboratory. *Experimental Infrastructure for Exploration of Exascale Computing (eX3)*. https://www.ex3.simula.no/publications. The research presented in this paper has benefited from the Experimental Infrastructure for Exploration of Exascale Computing (eX3), which is financially supported by the Research Council of Norway under contract 270053. 2025.

[16] Haohan Ye et al. "IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models". In: *arXiv preprint arXiv:2308.06721* (2023).

[17] Hu Ye et al. "IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models.(2023)". In: *arXiv preprint arXiv:2308.06721* (2023).

[18] Chenshuang Zhang et al. "Text-to-image diffusion models in generative ai: A survey". In: *arXiv preprint arXiv:2303.07909* (2023).

[19] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 3836–3847.

[20] Tianyi Zhang et al. "A survey of diffusion based image generation models: Issues and their solutions". In: *arXiv preprint arXiv:2308.13142* (2023).

[21] Qilong Zhangli et al. "SceneTextGen: layout-agnostic scene text image synthesis with diffusion models". In: *arXiv e-prints* (2024), arXiv–2406.

[22] Mingkang Zhu et al. "LogoSticker: Inserting Logos Into Diffusion Models for Customized Generation". In: *European Conference on Computer Vision*. Springer. 2024, pp. 363–378.

VIII. Individual contribution

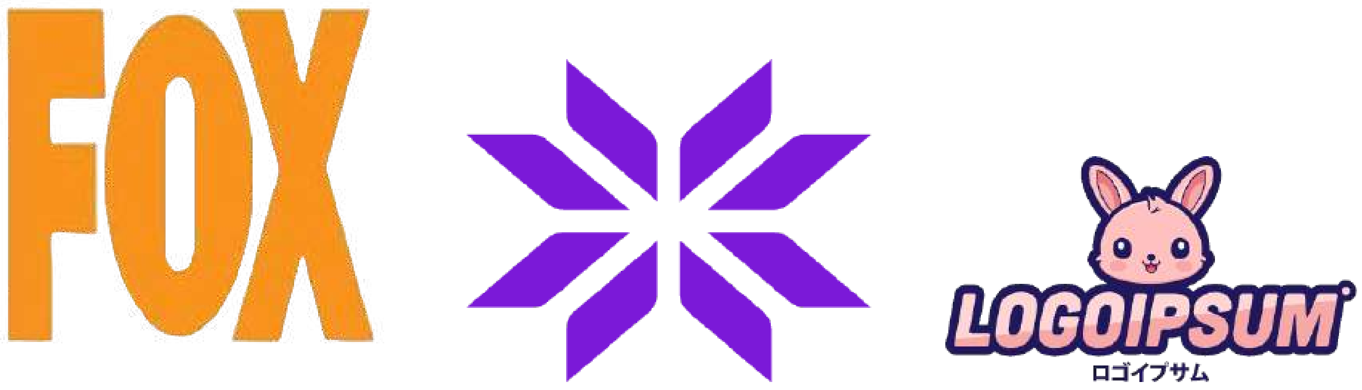
Brage Dybesland Ramberg: Code: (100%) Report: (40%)

Brage served as the primary developer for the code in this project. He was responsible infrastructure, environmental setup, and experimentation on the SIMULA eX3 cluster, with help from the two other members. Brage also conducted visual inspection of generated outputs, tuned parameters, proposed strategies for improvements. Additionally, he drafted most of the rapport sections about experimentation and results section.

Vaskar Shrestha: Code: 40% Report: 100% Vaskar contributed primarily to the documentation in this project. He was responsible for drafting key sections of the report, including the Abstract, Literature Review, Conclusion, and Acknowledgements. In addition, he also participated in the technical implementation by assisting with environment setup and running selected versions of the code on the SIMULA eX3 cluster.

Dung participates in both the technical development and the reporting aspects of the project. She was primarily responsible for implementing the base approach of single-modal and finding ideas, contributing to the core codebase and laying the groundwork for further experimentation. On the documentation side, Dung authored the Methodology and Discussion sections, providing a clear explanation of the technical process and a critical analysis of the results. She also assisted in reviewing and refining other parts of the report to ensure coherence and quality.

Appendix: Logo Inputs and Image Collages on seed 42



FOX Logo

Purple Star Logo

Rabbit Logo

Fig. 6: Logos used as IP-Adapter input for guiding generation. Each logo is used across multiple prompt scenes and grid configurations.

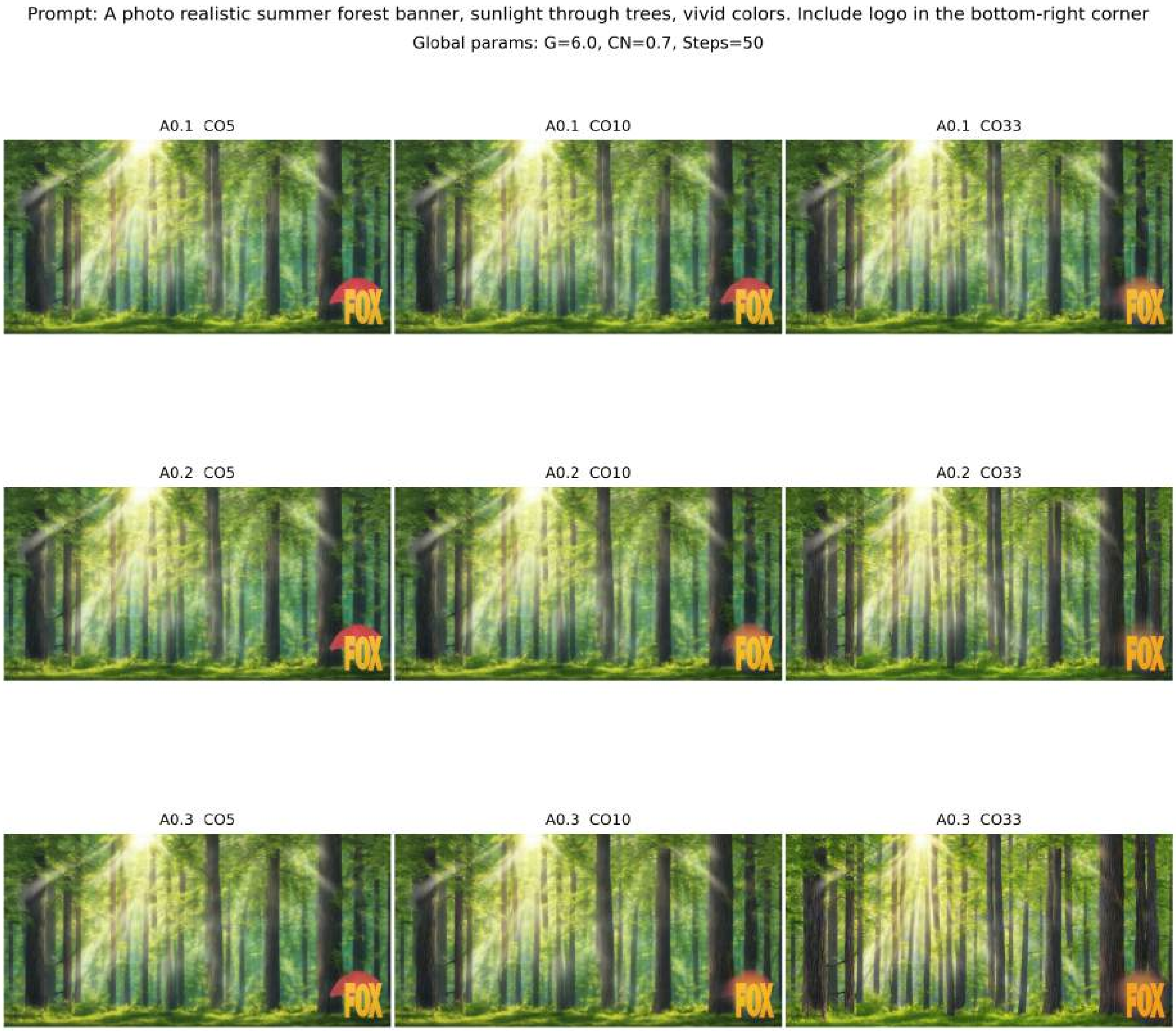
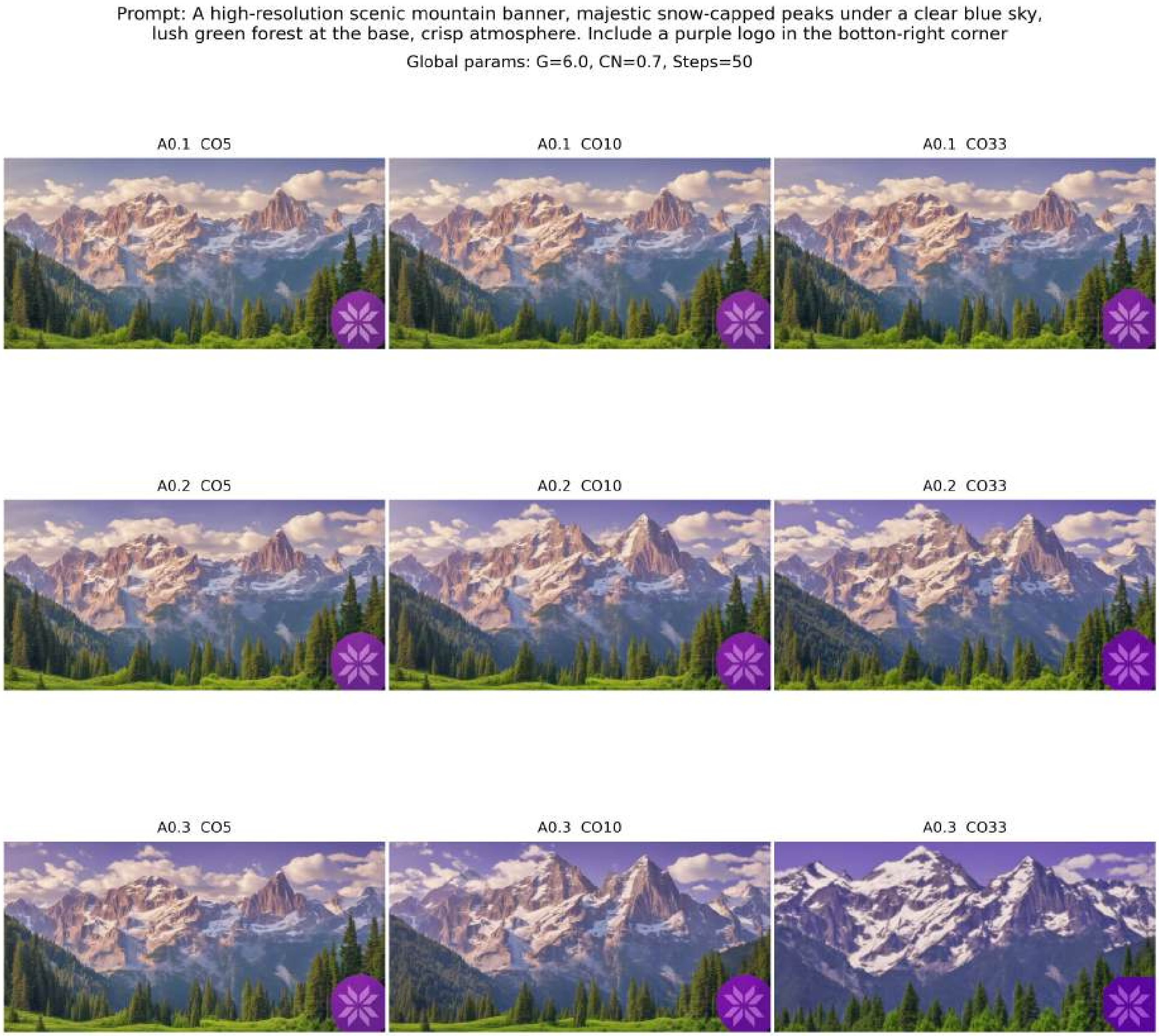

Prompt: A photo realistic summer forest banner, sunlight through trees, vivid colors. Include logo in the bottom-right corner
Global params: G=6.0, CN=0.7, Steps=50
A0.1 CO5
A0.1 CO10
A0.1 CO33
A0.2 CO5
A0.2 CO10
A0.2 CO33
A0.3 CO5
A0.3 CO10
A0.3 CO33
FOX

Fig. 7: Collage for the ”Forest scene” prompt with focus on adapter (A) and cutoff (CO) values.



Prompt: A high-resolution scenic mountain banner, majestic snow-capped peaks under a clear blue sky,
lush green forest at the base, crisp atmosphere. Include a purple logo in the botton-right corner

Global params: G=6.0, CN=0.7, Steps=50

A0.1 CO5

A0.1 CO10

A0.1 CO33

A0.2 CO5

A0.2 CO10

A0.2 CO33

A0.3 CO5

A0.3 CO10

A0.3 CO33

Fig. 8: Full-resolution collage for the ”Mountain scene” prompt with focus on adapter (A) and cutoff (CO) values.

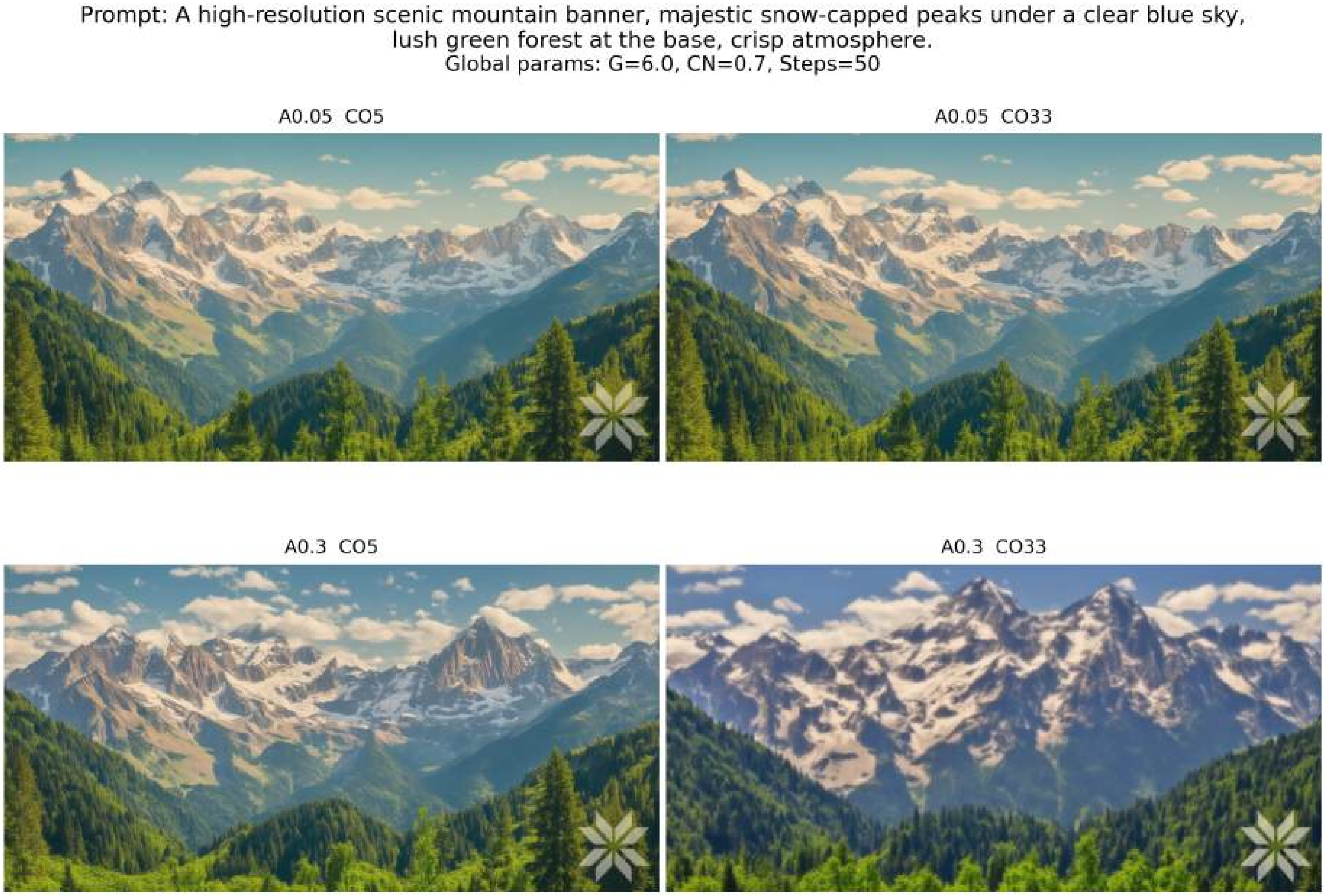

Prompt: A high-resolution scenic mountain banner, majestic snow-capped peaks under a clear blue sky,
lush green forest at the base, crisp atmosphere.
Global params: G=6.0, CN=0.7, Steps=50
A0.05 CO5
A0.05 CO33
A0.3 CO5
A0.3 CO33

Fig. 9: Collage for the ”Mountain scene” prompt with focus on adapter (A) and cutoff (CO) values.

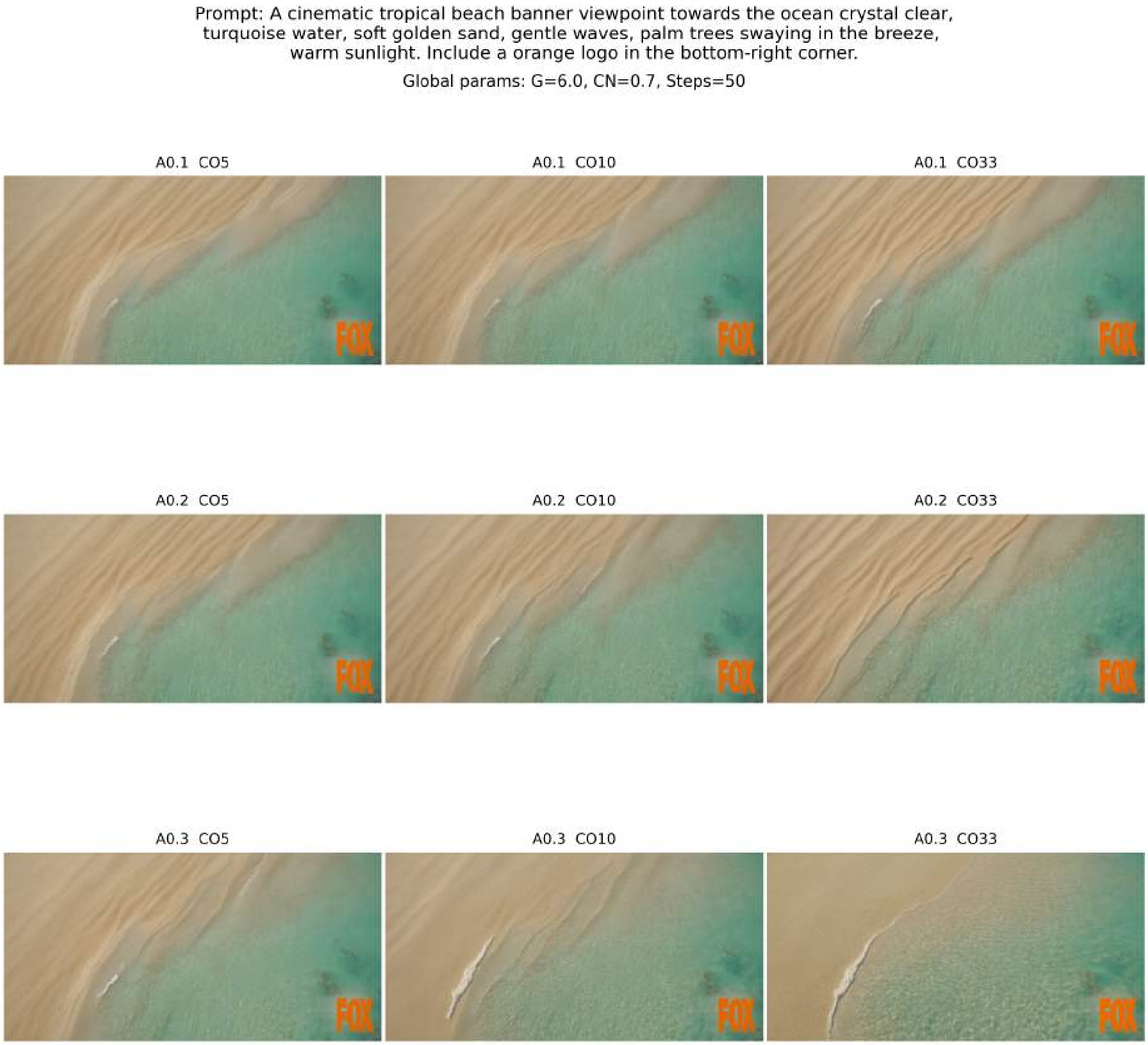
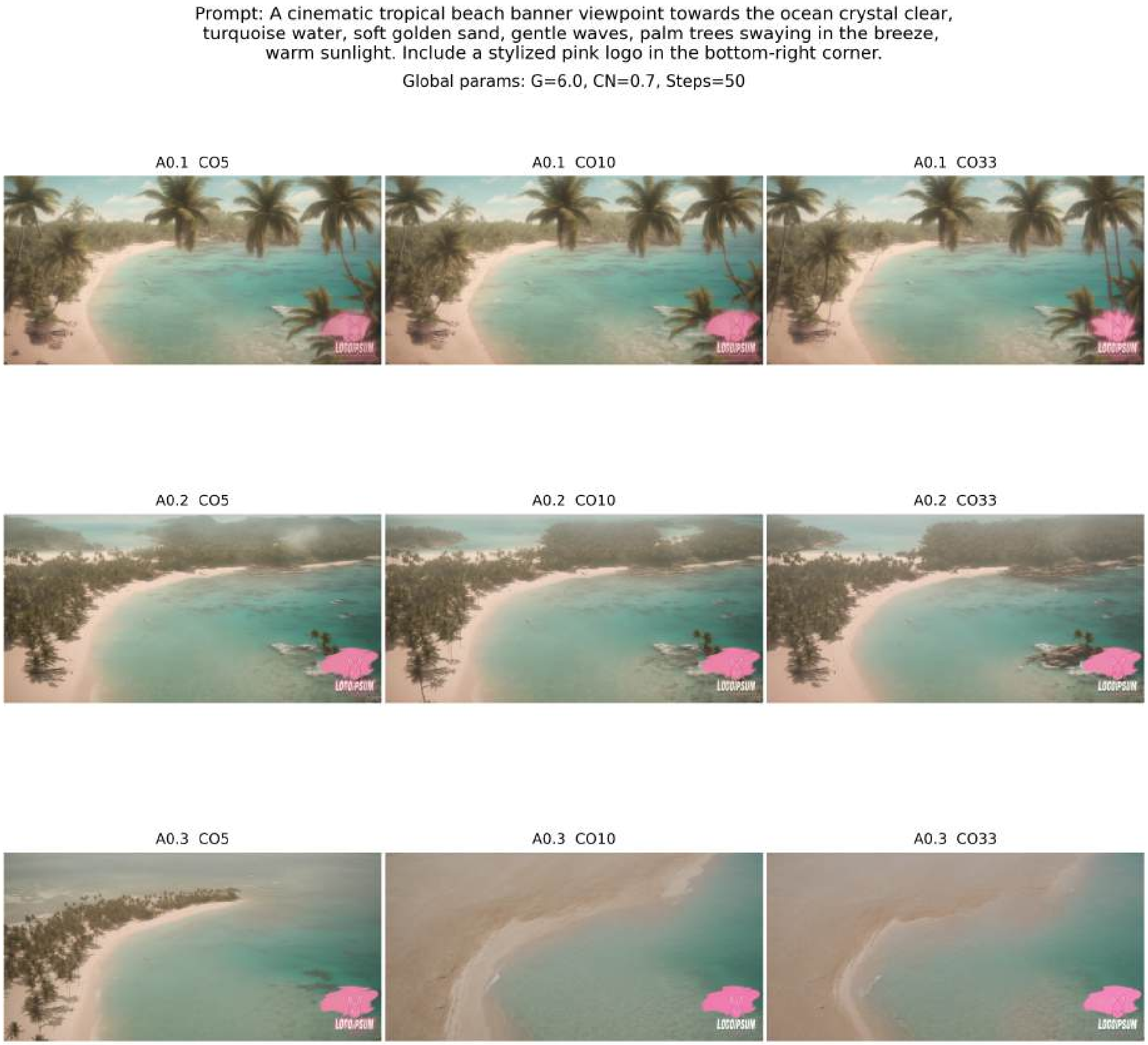

Prompt: A cinematic tropical beach banner viewpoint towards the ocean crystal clear,
turquoise water, soft golden sand, gentle waves, palm trees swaying in the breeze,
warm sunlight. Include a orange logo in the bottom-right corner.
Global params: G=6.0, CN=0.7, Steps=50
A0.1 CO5
A0.1 CO10
A0.1 CO33
A0.2 CO5
A0.2 CO10
A0.2 CO33
A0.3 CO5
A0.3 CO10
A0.3 CO33

Fig. 10: Collage for the ”Beach scene” prompt with focus on adapter (A) and cutoff (CO) values.



Prompt: A cinematic tropical beach banner viewpoint towards the ocean crystal clear, turquoise water, soft golden sand, gentle waves, palm trees swaying in the breeze, warm sunlight. Include a stylized pink logo in the bottom-right corner.

Global params: G=6.0, CN=0.7, Steps=50

Fig. 11: Collage for the ”Beach scene 2” prompt with focus on adapter (A) and cutoff (CO) values.

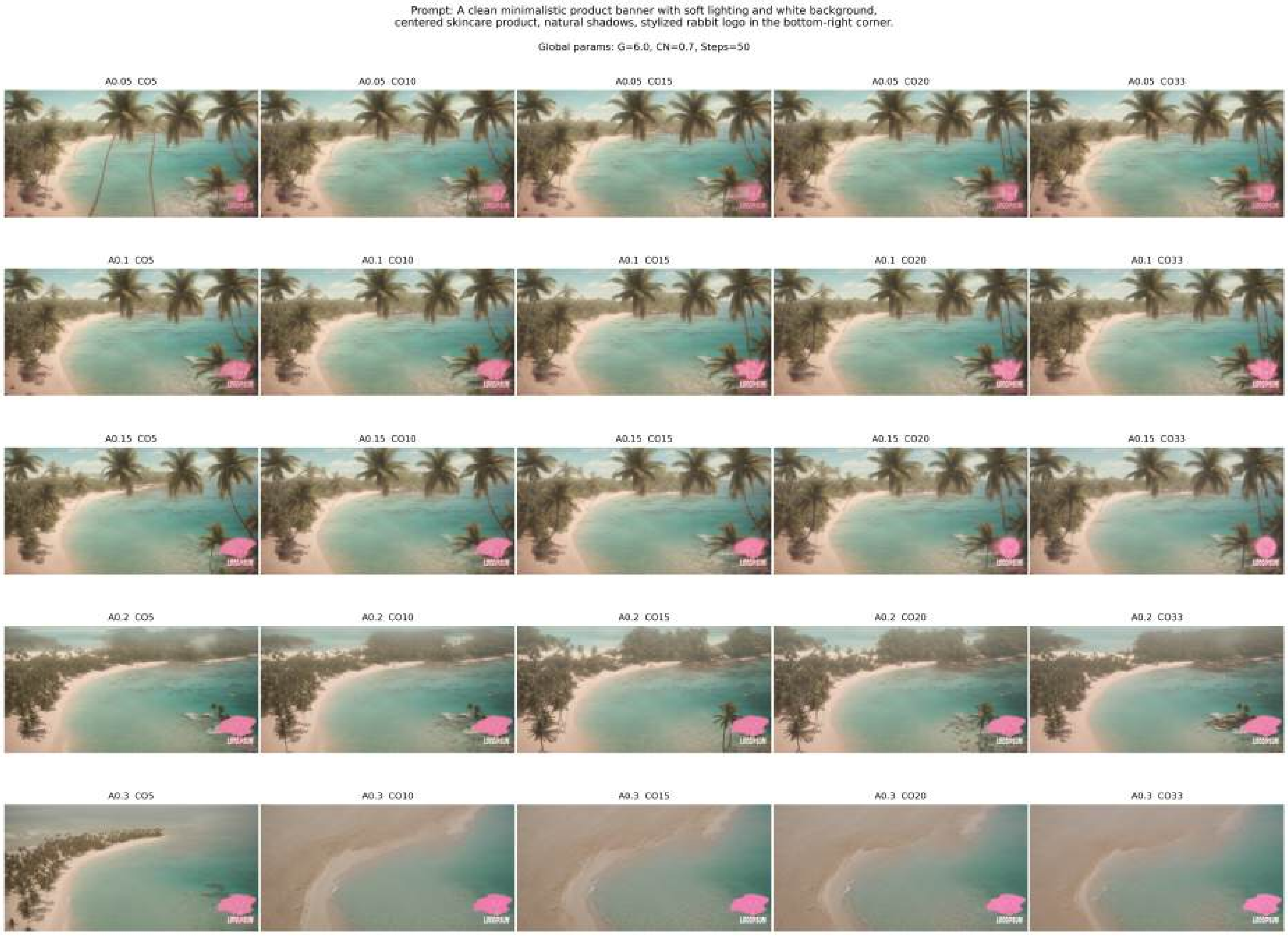


Fig. 12: Collage for the ”Beach scene 2” prompt with focus on adapter (A) and cutoff (CO) values.